Article ■

Effects of Data Anonymization by Cell Suppression on Descriptive Statistics and Predictive Modeling Performance

Lucila Ohno-Machado, MD, PhD, Staal Vinterbo, PhD, Stephan Dreiseitl, PhD

Abstract Protecting individual data in disclosed databases is essential. Data anonymization strategies can produce table ambiguation by suppression of selected cells. Using table ambiguation, different degrees of anonymization can be achieved, depending on the number of individuals that a particular case must become indistinguishable from. This number defines the level of anonymization. Anonymization by cell suppression does not necessarily prevent inferences from being made from the disclosed data. Preventing inferences may be important to preserve confidentiality. We show that anonymized data sets can preserve descriptive characteristics of the data, but might also be used for making inferences on particular individuals, which is a feature that may not be desirable. The degradation of predictive performance is directly proportional to the degree of anonymity. As an example, we report the effect of anonymization on the predictive performance of a model constructed to estimate the probability of disease given clinical findings.

■ **J Am Med Inform Assoc.** 2002; 9(Nov-Dec suppl):S115–S119. DOI 10.1197/jamia.M1241.

Introduction

The protection of privacy in health care information has received increasing attention from legislators, health care provider organizations, and consumers.[1] Legal and ethical issues arise from the increasing transmission of health care data over the Internet, and the utilization of such data for a variety of purposes.[2,3] In particular, data collected for provision of primary care can be re-utilized for epidemiological characterizations, and for the construction of predictive models.

The definition of data anonymization is currently not very specific. Several authors consider "de-identified" data sets those in which unique identifiers such as SSN or other set of attributes have been removed. Sweeney[4] has shown that this is not sufficient to hinder identification, as other publicly available data sets can be used to link information and uniquely identify individuals.

Several anonymization strategies exist (see [5,6] for a review). Some of the anonymization strategies are based on cell suppression and can produce an objective metric of anonymization.[6,7] A very simple example is given in Figure 1. In this example, every row is made indistinguishable from one other one by suppression of two cells. The averages for the columns are preserved, but the unique identification of a row is not possible (i.e., the table is ambiguated or anonymized at level = 2).

Certain anonymization strategies take into consideration which attributes should be protected ("relative anonymization"). Evidently the easiest way to protect that information is to completely delete the

Affiliations of the authors: Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, Division of Health Sciences and Technology, Massachusetts Institute of Technology, Boston, Massachusetts (LO-M, SV); and Department of Software Engineering for Medicine, Polytechnic University of Upper Austria, Hagenberg, Austria (SD). e-mail: <machado@dsg.harvard.edu>.

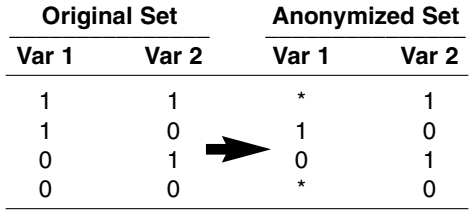

Figure 1 Simple example of how cell suppression (denoted *) makes one row (record) indistinguishable from at least one other row. Unique identification of the record is not possible. Column averages are preserved.

attribute from the disclosed dataset. In cases where the dataset is disclosed to be used for predictive modeling, this deletion may make the data useless. Furthermore, intermediate degrees of deletions (or cell suppressions) may be sufficient for protecting privacy and yet still result in data sets that are useful for predictive modeling. The relationship between degree of anonymization and predictive modeling capability has not been fully investigated. It is this issue that we address in our experiment.

Materials and Methods

We have chosen a real clinical data set of moderate size to test the hypothesis that anonymized tables can preserve certain characteristics of the data and still be useful for predictive modeling.

Data

We used a data set of 250 patients *(training sample)* suspected of having myocardial infarction (MI) at admission to the emergency room. A set of 700 cases from a different hospital was used for validation (*test set*). This data set was used previously in other predictive modeling studies[8-11] and was chosen because we knew that it could perform well when the whole data were present, and that few variables were sufficient to model the classification problem. This data set contained 46 variables representing clinical findings related to MI.

Methods

Data from the training set were anonymized using table ambiguation by cell suppression, as described in [7]. The test set was used to estimate predictive performance on a set of previously unseen cases. We constructed one logistic regression model per degree of anonymity, and assessed the classification performance using areas under the ROC curves. Cases in which the outcome variable was suppressed were eliminated. Data were imputed in suppressed cells by substituting the missing value by the average of the remaining values for non-suppressed cells in the same column.

Table 1 ■

Areas Under the ROC Curve Corresponding to Different Levels of Anonymity

Anonymity	AUC	Std. Dev.	N	Supp.
Training	0.985	0.011	250	0
Test set	0.875	0.016	700	0
2	0.782	0.020	226	1158
3	0.820	0.018	222	1471
4	0.759	0.021	217	1690
5	0.762	0.021	216	1835
6	0.758	0.020	211	1956
7	0.764	0.020	212	2055
8	0.742	0.021	211	2159
9	0.766	0.020	209	2228
10	0.777	0.019	208	2288
20	0.743	0.021	200	2713
30	0.688	0.022	190	2958
40	0.650	0.021	180	3161
50	0.687	0.022	180	3300
60	0.615	0.023	171	3411
70	0.702	0.023	160	3513
80	0.654	0.022	159	3590
150	N/A	N/A	126	3878

AUC = Area Under the ROC Curve
N = number of cases with non-missing values for the dependent variable
Supp. = number of suppressed cells

We used 18 data files with varying number of cell suppressions. The suppression algorithm allows specification of the "bin size." The "bin size" denotes the number of cases from which any particular case in the data set is made undistinguishable. For example, a "bin size" or anonymization level of 2 indicates that each case is undistinguishable from at least one other in the data set, and a "bin size" of 150 indicates that each case is undistinguishable from at least 150 others.

We used the SAS PROC LOGISTIC[12] with default parameters to build 18 different logistic regression models, which were evaluated in a set of previously unseen cases. The "bin size," number of cases, and number of overall suppressions are given in Table 1. We calculated descriptive statistics with PROC MEANS in SAS.

Results

Table 1 lists the areas under the ROC curve and their respective standard deviations when the model is applied to the test set. These results can be visually inspected in Figure 2. Figure 3 shows the effect of cell suppressions for the attribute "age." Figure 4 shows the effects of cell suppression on the proportion of cases in which "Smoker" = 1.

The deterioration of performance in the test set (in comparison with the baseline model) for all "bin sizes" is statistically significant for $\alpha = 0.05$. Even suppressions required for small "bin sizes" of 2 and 3 resulted in significant reduction in the predictive ability of these models. Conversely, the areas under the ROC curve for the training set increased very fast, achieving a perfect index of 1 for a "bin size" of only 3, demonstrating how fast the model was overfitting the data. The goodness-of-fit Hosmer-Lemeshow statistics increased accordingly. Although significantly different, data resulting from anonymization at levels 2 to 20 still produced reasonable areas under the curve, indicating that predictive models of MI could be built. The models were clearly not good for "bin sizes" greater than 20.

Discussion

The logistic regression model was used to verify the influence of the number of cell suppressions (a proxy for anonymity, as explained in the introduction) on predictive model performance. Several other classes of models could be used, such as neural networks, classification trees, rough sets, and support vector machines, to name just a few. Our previous work with this data set did not indicate that more sophisticated machine learning models such as neural networks or rough sets could yield significantly better performances (at $\alpha = 0.05$) in this data set when no missing values were present. We therefore did not expect that this would be the case in this experiment as well, especially given the fact that we did not eliminate cases with missing values, but rather imputed the data whenever necessary.

In this experiment, we used a very simple imputation method. Imputation was important because if we had instead decided to eliminate cases with missing values, model construction would not be possible

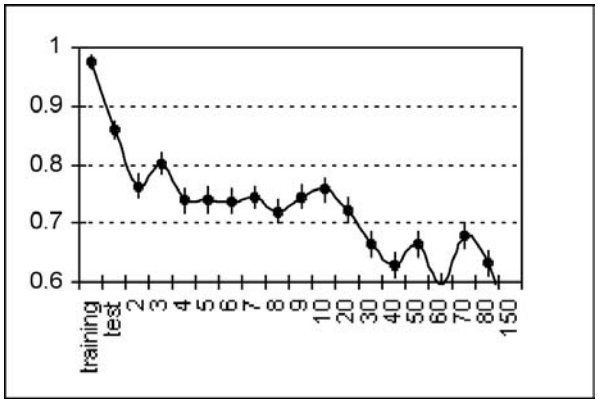

Figure 2 Areas under the ROC by degree of anonymization.

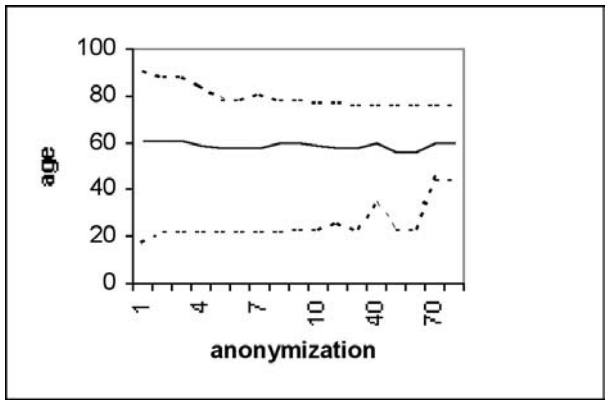

Figure 3 Max, mean, and min(age) for different levels of anonymization. Note that the means for age was preserved throughout the process. Extreme values were increasingly removed, since the age distribution was approximately normally distributed (i.e., values at the tails of the normal curve were removed).

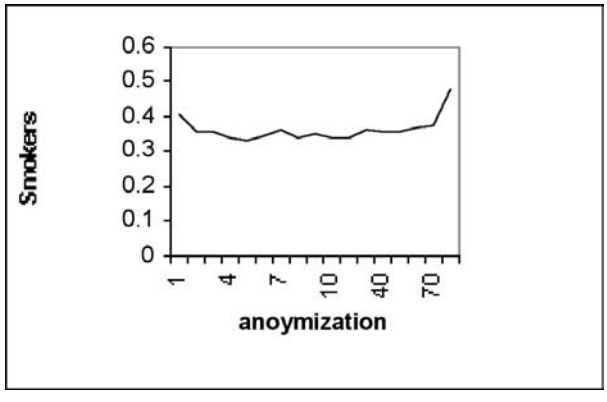

Figure 4 Proportion of smokers for different levels of anonymization. The proportion of cases does not vary significantly.

beyond very small "bin sizes" because of the scarcity of training cases. We believe that more sophisticated imputation methods would result in slightly better performances.

Another possibility for predictive model construction in the presence of missing values would be to use algorithms that can ignore this missing data. This is not the case for the algorithms mentioned above. However, other algorithms such as nearest-neighbor methods do not require all fields in an entry to be given, but can calculate distances based on the entries present. We did not pursue this approach here, because the discriminatory power of these methods is generally not as good as that of more sophisticated algorithms.

The algorithm that we used for cell suppression was very simple and used heuristics to reduce the complex search for optimal suppressions to a manageable level. In our limited experience with this algorithm when applied to small data sets, the number of suppressions required were not dramatically different from the number that could be obtained by doing an exhaustive search. Compared to a similar algorithm described by Orhn,[6] we achieved fewer suppressions.

It can be seen from Table 1 that several cases had to be removed, as the suppression occurred in the dependent variable MI. In case the outcome of interest can be anticipated (e.g., we want to release the data for researchers trying to predict the diagnosis of myocardial infarction), then it might be reasonable to limit the suppressions to other attributes as much as possible. This way, the loss of a significant number of cases could be avoided, although the overall number of total suppressed cells may be higher. It remains to be seen whether the decrease in performance was related to the number of cases alone. Related experiments in which we randomly removed cases from a similar data set indicated a high degree of data redundancy.[13] The fact that the performance decreased significantly when as few as 24 cases were removed seems to indicate that the anonymization procedure was the main factor for the performance decay. More detailed experiments in this area are necessary to verify this hypothesis.

We demonstrated that, in this data set, an anonymity level of 2 would already result in predictive performance that would be significantly different that that of the baseline. The question remains whether this would be an "anonymous enough" data set, and whether the reduction in predictive performance is acceptable. It should be noted that a data set with an anonymity level of 10 would result in approximately the same predictive performance.

The choice of the correct level of anonymization is not purely analytical. It is very dependent on the expected use of the data set, and it would be advisable to construct a decision-theoretic model that included an accurate utility set to determine the "right" levels of anonymization for different users. An additional consideration in the choice of anonymization level has to be the fact that cell suppression at level 2 may be partially reverted by "guessing" the missing values and then checking whether the complete table obtained this way could possibly lead to the cell suppressions in the original anonymized table. While this approach is computationally demanding, as all possible combinations of guessed values have to be checked, it still shows that higher levels of anonymity may be required to counter such disambiguation efforts.

Conclusion

Preserving an individual's privacy in disclosed data sets has become an important concern now that large amounts of clinical and genomic data can be transferred worldwide using the Internet. Anonymization of data sets is important to preserve privacy, and can be achieved by making cases indistinguishable from a pre-determined number of other cases in the same data set. One of the ways to achieve this is to selectively suppress cells from a table. If privacy is to be protected, hindering undesirable inferences in the disclosed data set has to be taken into consideration.

We show that predictive models can be constructed from "anonymized" data sets. The results reported here can be seen as a first step into this research direction, and we will continue to study the impact of anonymization algorithms on predictive modeling. We have analyzed a series of data sets with varying degrees of anonymization and demonstrated the impact of anonymization level on predictive performance. Further work will be required to define acceptable tradeoffs between predictive performance and anonymization levels.

This work was funded in part by grant R01-LM653801 from the National Library of Medicine. We thank Dr. Hamish Fraser for providing the data set for this experiment.

Reprinted from the Proceedings of the 2001 AMIA Annual Symposium, with permission.

References ■

1. Rules and Regulations. Federal Register 65(250), Dec 28, 2000.
2. Buckovich SA, Rippen HE, Rozen, MJ. Driving Toward Guiding Principles: A Goal for Privacy, Confidentiality, and Security of Health Information. J Am Med Inform Assoc 1999;6: 122-133.
3. Campbell SG, Gibby GL, Collingwood S. The Internet and electronic transmission of medical records. J Clin Monit. 1997;13:325-34.
4. Sweeney L. Guaranteeing anonymity when sharing medical data, the DataFly system. In: Masys D R, editor. Proc. AMIA Fall Symposium. 1997:51-5.
5. Fischetti M, Salazar JJ. Models and Algorithms for the 2-Dimensional Cell Suppression Problem in Statistical Disclosure Control. Mathematical Programming. 1999:283-312.
6. Øhrn A, Ohno-Machado L. Using Boolean Reasoning to Anonymize Databases. Artificial Intelligence in Medicine 1999;15:235-254.
7. Vinterbo S, Ohno-Machado L. Table Disambiguation Via Cell Suppression. DSG Technical Report. Brigham and Women's Hospital.
8. Wang S, Ohno-Machado L, Fraser H, Kennedy L. Using Patient-Reportable Clinical History Factors to Predict Myocardial Infarction. Computers in Biology and Medicine 2001.
9. Vinterbo S, Ohno-Machado L. A Genetic Algorithm to Select Variables in Logistic Regression: Example in the Domain of Myocardial Infarction. Proc. AMIA Fall Symp. 1999; 984-8.
10. Dreiseitl S, Ohno-Machado L, Vinterbo S. Evaluating Variable Selection Methods for Diagnosis of Myocardial Infarction. Proc. AMIA Fall Symp. 1999; 246-50.
11. Kennedy RL, Burton AM, Fraser HS, McStay LN, Harrison RF. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: derivation and evaluation of logistic regression models. Eur Heart J. 1996;17:1181-91.
12. SAS Institute. SAS/STAT Version 8.1.
13. Ohno-Machado L, Fraser HS, Øhrn A. Improving Machine Learning Performance by Removing Redundant Cases in Medical Data Sets. Proc. AMIA Fall Symp. 1998; 523-527.